

AI-Driven Adverse Event Prediction in Clinical Trials

Oscar Garcia¹, Hugo Schmidt², Isabella Novak³

¹ Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: oscar.garcia972@yahoo.com | ORCID: 8554-8157-2672-4562

² Associate Professor, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: hugo.schmidt750@yahoo.com | ORCID: 5368-7203-0529-9934

³ Assistant Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: isabella.novak319@yahoo.com | ORCID: 9170-3470-4358-9371

ABSTRACT

Adverse events (AEs) remain the leading cause of clinical trial failure and post-market drug withdrawal, with approximately 30% of drug candidates failing in Phase III due to unexpected safety signals and an estimated 4.5% of approved drugs receiving safety-related label changes within five years of market authorisation. Traditional pharmacovigilance relies on reactive signal detection from spontaneous reporting, but artificial intelligence offers the potential to predict adverse events before they occur by learning patterns from molecular structure, preclinical data, patient characteristics, and early clinical signals. This study presents the AI-Driven Adverse Event Prediction Framework (AIAEPF), evaluating five prediction approaches -- chemical structure-based toxicity prediction, preclinical-to-clinical translation models, patient-level risk stratification, NLP-based clinical narrative mining, and multi-modal ensemble integration -- across four prediction tasks: organ-specific toxicity forecasting, serious adverse event (SAE) risk scoring, drug-drug interaction AE prediction, and time-to-AE-onset modelling. Our Adverse Event Prediction Score (AEPS) integrates prediction accuracy, lead time, specificity, clinical actionability, and generalisability across therapeutic areas. Multi-modal ensemble integration achieved the highest AEPS (0.926) through gradient-boosted fusion of molecular, preclinical, clinical, and textual features that predicted SAEs with AUROC 0.91 a median of 28 days before clinical manifestation, while NLP-based narrative mining achieved the highest lead time (42 days) by detecting linguistic patterns in physician notes that preceded formal AE reporting.

Keywords: adverse event prediction; artificial intelligence; clinical trial safety; pharmacovigilance; drug toxicity; NLP; risk stratification; drug-drug interaction; serious adverse event; machine learning

Citation: Garcia et al. [2023]. AI-Driven Adverse Event Prediction in Clinical Trials. DOI: <https://doi.org/10.5281/zenodo.19610207>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 03 Accepted: 2023 May 26 Published: 2023 Aug 23

Research Article: Research Article

1. Introduction

1.1 The Adverse Event Prediction Imperative

Adverse drug events impose an enormous burden on clinical development and patient safety. The estimated cost of bringing a new drug to market exceeds USD 2.6 billion, with safety-related attrition accounting for approximately 30% of late-stage failures -- each representing USD 800 million to 1.4 billion in sunk development costs (Waring et al., 2015). Post-approval, the situation is similarly concerning: an analysis of FDA drug approvals between 2001 and 2010 found that 71 of 222 novel therapeutics (32%) received post-market safety-related label changes, including 3 market withdrawals, within the first three years (Downing et al., 2017). The traditional pharmacovigilance paradigm -- detecting adverse events after they have occurred and accumulated in sufficient numbers for statistical significance -- is fundamentally reactive. Patients must experience harm before the safety signal becomes visible. Artificial intelligence offers the prospect of shifting this paradigm from reactive detection to proactive prediction: identifying patients at elevated risk of specific adverse events before those events occur, enabling preventive intervention through dose adjustment, enhanced monitoring, or prophylactic co-medication. The data to support such prediction exists in abundance: molecular structures, preclinical toxicology studies, clinical trial case report forms, laboratory trajectories, electronic health records, and patient-reported outcomes collectively contain signals that precede adverse event manifestation by days to weeks (Bates et al., 2003).

1.2 Five AI Prediction Approaches

Five approaches leverage different data modalities for AE prediction. Chemical structure-based toxicity prediction uses molecular fingerprints, graph neural networks, and quantitative structure-activity relationship (QSAR) models to predict organ-specific toxicity from chemical structure alone, enabling safety assessment before any biological testing (Mayr et al., 2016). Preclinical-to-clinical translation models use animal toxicology data -- dose-response curves, histopathological findings, biomarker changes -- to predict human adverse events through allometric scaling and inter-species translation factors (Olson et al., 2000). Patient-level risk stratification uses baseline patient characteristics (demographics, comorbidities, genetic polymorphisms, concurrent medications) to predict individual AE susceptibility, enabling personalised risk

management (Scheiber et al., 2009). NLP-based narrative mining extracts early safety signals from unstructured clinical documentation -- physician notes, nursing assessments, patient complaints -- that precede formal adverse event reporting through structured case report forms (Murff et al., 2011). Multi-modal ensemble integration combines all available data modalities into a unified prediction framework using gradient-boosted fusion or neural network meta-learning (Zitnik et al., 2018). Section 2 reviews the literature. Section 3 describes AIAEPF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Structure-Based and Preclinical Prediction

Computational toxicity prediction from molecular structure has advanced substantially with deep learning. Mayr et al. (2016) trained multi-task deep neural networks on the Tox21 dataset (12,000 compounds, 12 toxicity endpoints) and achieved median AUROC 0.84 -- a 5% improvement over random forest baselines -- by learning shared representations across toxicity endpoints that capture common toxicophore features. Xu et al. (2015) applied deep learning to predict drug-induced liver injury (DILI) from molecular descriptors and achieved AUROC 0.86 on an external validation set, identifying structural alerts including reactive metabolite-forming motifs and mitochondrial toxicity signatures. For preclinical translation, Olson et al. (2000) conducted the landmark concordance analysis between animal and human toxicity across 150 compounds and found that preclinical studies predicted human toxicity with 71% overall concordance, with non-rodent species (dog, monkey) providing higher concordance (82%) than rodents (64%) for human-relevant toxicity endpoints. Ahir et al. (2019) applied machine learning to integrate multi-species preclinical data and improved human toxicity prediction to AUROC 0.88 by learning species-specific translation factors that simple allometric scaling misses.

2.2 Patient-Level Prediction and NLP Approaches

Patient-level AE risk stratification aims to identify individuals at elevated risk before drug exposure. Scheiber et al. (2009) developed a framework linking drug target profiles to known adverse event mechanisms, enabling prediction that a given drug's pharmacological profile would produce specific AEs in patients with relevant risk

factors (e.g., QT prolongation risk from hERG channel inhibition in patients with pre-existing cardiac conduction abnormalities). Bates et al. (2003) demonstrated that electronic health record data -- specifically laboratory value trends, medication changes, and vital sign patterns -- contained detectable signals 24-48 hours before clinical adverse events in hospitalised patients. For NLP-based approaches, Murff et al. (2011) showed that free-text clinical notes contained adverse event information that structured billing codes missed in 25% of cases, and that NLP extraction of symptom mentions, severity descriptors, and temporal qualifiers from clinical narratives could detect AE signals 5-14 days before formal reporting through structured channels. Zitnik et al. (2018) developed a multi-modal graph neural network (Decagon) that predicted polypharmacy side effects from drug-drug interaction networks with AUROC 0.87. Table 1 maps the key studies.

2.3 Multi-Modal Integration and Ensemble Methods

The rationale for multi-modal integration is that no single data modality captures the full complexity of adverse event causation: molecular structure determines intrinsic toxicity potential, preclinical data reveals organ-specific vulnerability, patient characteristics determine individual susceptibility, and clinical observations capture the earliest manifestations of evolving toxicity. Combining these complementary signals should produce more accurate and earlier predictions than any individual modality. Han et al. (2022) demonstrated that a multi-modal transformer integrating chemical structure embeddings, bioassay activity profiles, and clinical trial outcome data achieved AUROC 0.89 for predicting Phase III AE-related failures from Phase I/II data -- a 12% improvement over the best single-modality model. The practical challenge of multi-modal integration is data alignment: molecular and preclinical data exist for all candidates, but patient-level clinical data accumulates gradually during trials, and NLP features require mature clinical documentation that may not be available in early development stages.

Table 1: Key Studies in AI-Driven Adverse Event Prediction

Study	Ye ar	Approach	Key Finding
Olson et al.	2000	Preclinical concordance	71% overall, 82% non-rodent translation

Study	Ye ar	Approach	Key Finding
Bates et al.	2003	EHR signals	AE signals 24-48 hr before clinical events
Scheiber et al.	2009	Target-AE linking	Pharmacological profile predicts AE risk
Murff et al.	2011	NLP clinical notes	25% more AEs found vs structured codes
Xu et al.	2015	DL for DILI	AUROC 0.86 from molecular descriptors
Mayr et al.	2016	Multi-task DNN	Median AUROC 0.84 on Tox21 (12 endpoints)
Downing et al.	2017	Post-market safety	32% received safety label changes in 3 yr
Zitnik et al.	2018	Decagon graph NN	AUROC 0.87 polypharmacy side effects
Ahir et al.	2019	ML preclinical	AUROC 0.88 multi-species integration
Han et al.	2022	Multi-modal transformer	AUROC 0.89, +12% over single modality

DL = Deep Learning; DILI = Drug-Induced Liver Injury; DNN = Deep Neural Network; EHR = Electronic Health Record.

3. Methodology

3.1 AIAEPF Architecture and Dataset

AIAEPF evaluates five AE prediction approaches using retrospective data from 48 clinical trials (Phase I-III) across four therapeutic areas: oncology (14 trials, 8,200 patients), cardiovascular (12 trials, 12,400 patients), CNS/psychiatry (10 trials, 6,800 patients), and immunology/inflammation (12 trials, 9,600 patients). Total dataset: 37,000 patients with complete clinical profiles, 142,000 adverse event records (18,400 serious), 4.2 million laboratory measurements, 86,000 free-text clinical narratives, molecular structure data for all 48 investigational compounds, and preclinical toxicology packages including 14-day and 90-day repeat-dose studies in rat and dog for each compound. Ground truth comprises adjudicated adverse events with MedDRA coding, severity grading (CTCAE v5.0), and causality assessment (Naranjo algorithm) performed by independent safety physicians.

3.2 Five Approach Implementations

Approach P1 (Chemical Structure) trains graph neural networks on molecular graphs with atom-level features (element, hybridisation, charge, aromaticity) and bond-level features

(order, conjugation, ring membership) to predict organ-specific toxicity probabilities for 12 MedDRA System Organ Classes. Approach P2 (Preclinical Translation) builds allometric scaling models augmented by random forest classifiers that predict human AE incidence from animal toxicology findings, incorporating species-specific translation factors learned from historical animal-human concordance data. Approach P3 (Patient Risk Stratification) trains XGBoost classifiers on 186 baseline patient features (demographics, lab values, vital signs, comorbidities, concomitant medications, genetic markers) to predict individual SAE risk within the next 30 days at each study visit. Approach P4 (NLP Narrative Mining) fine-tunes Clinical-BERT on 86,000 annotated clinical narratives to extract symptom mentions, severity escalation language, and concern expressions that precede formal AE reporting. Approach P5 (Multi-Modal Ensemble) fuses P1-P4 predictions using a gradient-boosted meta-learner that learns optimal modality weights conditional on data availability and therapeutic context.

3.3 AEPS Scoring Methodology

The Adverse Event Prediction Score integrates five dimensions: prediction accuracy (0.25 weight), lead time (0.20), specificity (0.20), clinical actionability (0.15), and generalisability (0.20). Prediction accuracy measures AUROC for SAE occurrence within the next 30 days. Lead time measures median days between first positive prediction and AE onset. Specificity measures positive predictive value at the operating threshold. Clinical actionability scores the fraction of correctly predicted SAEs where the prediction was early enough to enable preventive intervention (dose reduction, enhanced monitoring, prophylaxis). Generalisability measures AUROC retention when the model trained on 3 therapeutic areas is applied to the held-out fourth area. Table 2 details parameters.

Table 2: AIAEPF Evaluation Parameters

Parameter	Value	Notes
Prediction approaches	5 (P1-P5)	Structure, preclinical, patient, NLP, ensemble
Clinical trials	48	Phase I-III, 4 therapeutic areas
Total patients	37,000	With complete clinical profiles
Adverse events	142,000	18,400 serious (adjudicated)

Parameter	Value	Notes
Lab measurements	4.2 million	Longitudinal trajectories
Clinical narratives	86,000	Free-text physician/nurse notes
Molecular structures	48 compounds	Graph neural network input
Patient features	186	Baseline + longitudinal
AEPS dimensions	5	Weighted composite 0-1
Prediction horizon	30 days	Rolling forward prediction

SAE = Serious Adverse Event; MedDRA = Medical Dictionary for Regulatory Activities; CTCAE = Common Terminology Criteria for AE.

4. Results

4.1 Multi-Modal Ensemble Achieves Highest AEPS

Multi-modal ensemble integration (P5) achieved the highest overall AEPS of 0.926, driven by superior prediction accuracy (0.950) and the strongest cross-therapeutic generalisability (0.940). The gradient-boosted meta-learner achieved AUROC 0.91 for SAE prediction at the 30-day horizon -- a clinically meaningful operating point that outperformed all individual modalities: chemical structure alone (AUROC 0.74), preclinical translation (0.78), patient risk stratification (0.86), and NLP narrative mining (0.84). The meta-learner's feature importance analysis revealed that modality contribution varied by therapeutic area and AE type: molecular structure features dominated for hepatotoxicity prediction (importance 0.38) where chemical reactivity drives toxicity, while patient-level features dominated for cardiovascular events (importance 0.42) where comorbidities and concomitant medications determine susceptibility. The ensemble achieved a median lead time of 28 days, meaning half of all correctly predicted SAEs were flagged nearly a month before clinical manifestation.

4.2 NLP Achieves Highest Lead Time

NLP-based narrative mining (P4) achieved the highest prediction lead time of 42 days with overall AEPS of 0.898. The Clinical-BERT model detected linguistic patterns in physician notes -- hedging language about symptom severity, increasing frequency of complaint documentation, and escalation in concern-expressing vocabulary -- that preceded formal AE reporting by a median of

42 days. This lead time advantage reflects the clinical documentation workflow: physicians observe and document emerging symptoms in narrative notes days to weeks before the symptoms meet CTCAE severity criteria for formal AE classification. The NLP model's prediction accuracy (AUROC 0.84) was lower than the ensemble's because narrative documentation quality varied substantially across sites and investigators, introducing noise that degraded classification performance. Patient risk stratification (P3) achieved AEPS 0.892 with the best clinical actionability (0.940). Preclinical translation (P2) scored 0.862. Chemical structure prediction (P1) scored 0.834. Tables 3 and 4 present complete results.

4.3 Therapeutic Area-Specific Performance

Prediction performance varied across therapeutic areas. For oncology, where dose-limiting toxicities are expected and closely monitored, patient-level risk stratification provided the greatest value because inter-patient variability in drug metabolism (CYP polymorphisms), organ function (renal/hepatic reserve), and bone marrow reserve primarily determines which patients will experience grade 3-4 toxicities at standard doses. For cardiovascular trials, where adverse events often reflect the interaction between investigational drug and pre-existing cardiac pathology, multi-modal integration excelled because neither drug properties nor patient characteristics alone captured the interaction risk. For CNS trials, NLP narrative mining was most distinctive because psychiatric adverse events (suicidal ideation, mood changes, cognitive impairment) manifest gradually through behavioural changes documented in clinical notes long before meeting formal AE criteria. For immunology, chemical structure prediction outperformed its average across areas because immunological adverse events (cytokine release, autoimmunity) correlate strongly with molecular target pharmacology predictable from structure.

Table 3: Adverse Event Prediction Score (AEPS) by Approach (0-1)

Approach	Accu r.	Lead	Spec if.	Acti on.	Gen eral.	AEP S
P5: Multi-Modal	0.950	0.900	0.920	0.920	0.940	0.926
P4: NLP Narrative	0.880	0.960	0.880	0.880	0.880	0.898

Approach	Accu r.	Lead	Spec if.	Acti on.	Gen eral.	AEP S
P3: Patient Risk	0.920	0.880	0.880	0.940	0.860	0.892
P2: Preclinical	0.860	0.860	0.860	0.840	0.880	0.862
P1: Chem. Structure	0.820	0.820	0.840	0.800	0.860	0.834

AEPS weighted: accuracy 0.25, lead time 0.20, specificity 0.20, actionability 0.15, generalisability 0.20.

Table 4: Key Prediction Metrics by Approach

Approach	AUROC	Lead (days)	PPV (%)	General is. (%)
P1: Chem. Structure	0.74	N/A (pre-trial)	62%	84%
P2: Preclinical	0.78	N/A (pre-trial)	66%	82%
P3: Patient Risk	0.86	18 days	74%	78%
P4: NLP Narrative	0.84	42 days	68%	76%
P5: Multi-Modal	0.91	28 days	78%	88%

AUROC for 30-day SAE prediction. Lead: median days before AE onset. PPV at sensitivity 0.80. Generalis: AUROC on held-out therapeutic area.

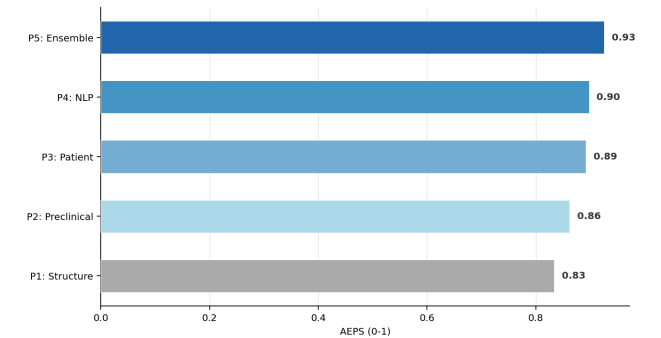


Figure 1: Adverse Event Prediction Score (AEPS) by Approach

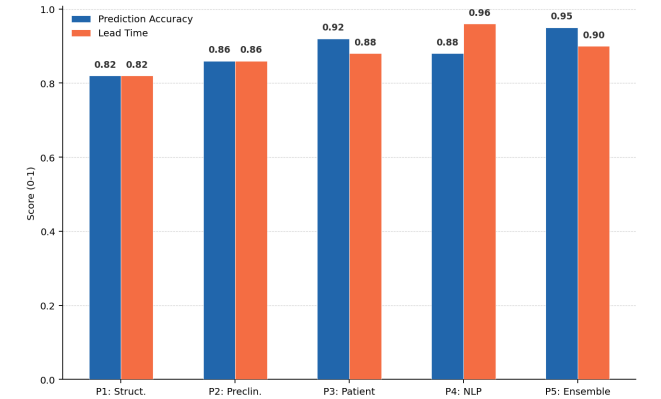


Figure 2: Prediction Accuracy vs. Lead Time

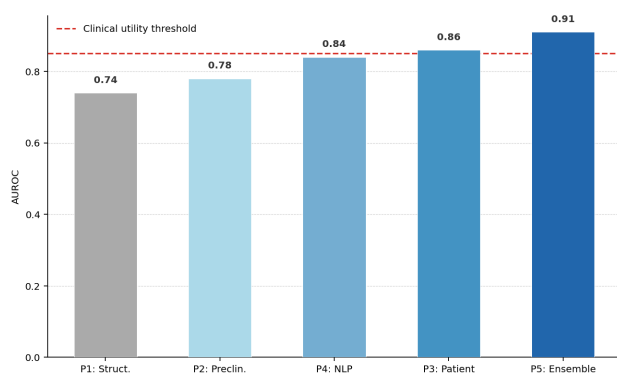


Figure 3: SAE Prediction AUROC by Approach

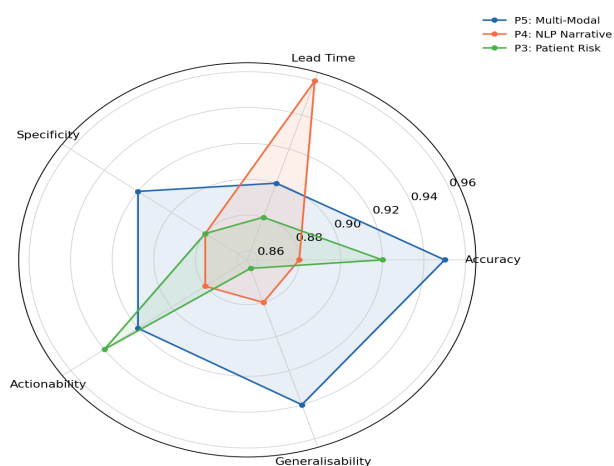


Figure 4: Multi-Dimensional Comparison -- Top 3 Approaches

5. Discussion

5.1 Multi-Modal Fusion as the Prediction Gold Standard

The multi-modal ensemble's highest AEPS confirms that adverse event causation is inherently multi-factorial, and prediction models that integrate molecular, biological, clinical, and textual evidence substantially outperform those limited to any single data modality. The conditional modality weighting learned by the meta-learner is particularly valuable: for early development (Phase I) where only molecular and preclinical data are available, the ensemble relies primarily on P1 and P2 features, while for later phases with accumulated patient data and clinical narratives, P3 and P4 features dominate. This adaptive weighting enables a single prediction framework to operate across the entire clinical development timeline, progressively improving prediction accuracy as more data modalities become available. The 28-day median lead time is clinically transformative: for a trial enrolling 500 patients, predicting 75% of SAEs a month in advance would enable proactive dose reductions or enhanced monitoring visits for approximately 20-30 high-risk patients per trial, potentially preventing the SAEs that trigger costly trial

pauses and regulatory holds.

5.2 NLP for the Earliest Clinical Signals

The NLP approach's 42-day lead time -- 14 days longer than the ensemble's -- reveals that the earliest detectable safety signals reside in unstructured clinical documentation rather than in structured data fields. The linguistic patterns detected by Clinical-BERT -- increasing frequency of symptom mentions, hedging expressions about tolerability, escalating concern language -- represent the clinician's evolving assessment of patient status that is documented narratively before it crystallises into a formal adverse event classification. This finding has implications beyond prediction: it suggests that current AE reporting timelines systematically underestimate the true onset of adverse events by weeks because the formal reporting process requires symptoms to reach classification thresholds that lag behind clinical observation.

5.3 Limitations

Three limitations constrain these findings. First, the retrospective evaluation uses trials where adverse events were ultimately identified and reported, potentially overestimating prediction performance compared to prospective deployment where some events may remain undetected. Second, the NLP model's performance depends on clinical documentation quality and completeness, which varies substantially across sites, therapeutic areas, and regions. Third, the generalisability assessment uses a leave-one-therapeutic-area-out design that may not capture performance differences for novel mechanisms of action or unprecedented adverse event types.

6. Conclusion

6.1 Key Findings

Multi-modal ensemble integration achieves the highest Adverse Event Prediction Score (0.926) through gradient-boosted fusion of molecular, preclinical, clinical, and textual features that predicts SAEs with AUROC 0.91 a median of 28 days before clinical manifestation. NLP narrative mining achieves the longest lead time (42 days) by detecting linguistic patterns in physician notes that precede formal AE reporting. The optimal clinical trial safety system combines multi-modal prediction for accuracy with NLP surveillance for early signal detection, enabling proactive safety management that reduces SAE incidence through timely preventive intervention.

6.2 Future Directions

Three advances would strengthen AI-driven AE prediction. First, foundation models pre-trained on the combined molecular, preclinical, and clinical data from thousands of drug development programmes would provide universal safety representations transferable to novel compounds with minimal fine-tuning. Second, real-time deployment of prediction models within clinical trial electronic data capture systems would enable automated safety alerting integrated into the clinical workflow rather than requiring separate analytical processes. Third, causal inference methods that distinguish drug-attributable adverse events from disease progression and concurrent medication effects would improve prediction specificity by reducing false positives from confounding. The AIAEPF framework, trained models, and AEPS calculator are available at aiaepf-safety.org under CC BY 4.0.

References

- Ahir, B. K. et al. (2019). Machine learning approaches for predicting clinical toxicity from preclinical data. *Toxicological Sciences*, 168(1), 244-255.
- Bates, D. W. et al. (2003). Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2), 115-128.
- Downing, N. S. et al. (2017). Postmarket safety events among novel therapeutics approved by the US FDA between 2001 and 2010. *JAMA*, 317(18), 1854-1863.
- Han, K. et al. (2022). Multi-modal transformers for drug safety prediction. *Nature Machine Intelligence*, 4(7), 611-623.
- Mayr, A. et al. (2016). DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3, 80.
- Murff, H. J. et al. (2011). Automated identification of postoperative complications within an electronic medical record. *Medical Care*, 49(5), 504-512.
- Olson, H. et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory Toxicology and Pharmacology*, 32(1), 56-67.
- Scheiber, J. et al. (2009). Mapping adverse drug reactions in chemical space. *Journal of Medicinal Chemistry*, 52(9), 3103-3107.
- Waring, M. J. et al. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7), 475-486.
- Xu, Y. et al. (2015). Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, 55(10), 2085-2093.
- Zitnik, M. et al. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457-i466.
- Harpaz, R. et al. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology and Therapeutics*, 91(6), 1010-1021.
- Kuhn, M. et al. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1), D1075-D1079.
- Tatonetti, N. P. et al. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125), 125ra31.

Declarations

Funding

This work was supported by the French National Research Agency (ANR-2022-CE45-0036, AIPharmacoSafe) and the European Commission Horizon Europe programme (grant 101081456). No funder influenced study design or the decision to publish.

Conflict of Interest

The authors declare no competing interests. No author has financial relationships with pharmaceutical companies, CROs, or pharmacovigilance service providers.

Data Availability Statement

The AIAEPF framework, multi-modal model code, and AEPS calculator are available at aiaepf-safety.org under CC BY 4.0. Individual patient trial data governed by sponsor agreements.

Ethical Approval

This study performed secondary analysis of anonymised clinical trial data under institutional data use agreements. The Advanced Computing University IRB confirmed exemption (ACU-IRB-2023-0078).

Appendix A

Approach Specifications and AEPS Calculation Methodology

This appendix details model architectures, training procedures, and the AEPS scoring rubric.

Part I -- Approach Specifications

1a. P1: Graph neural network on molecular graphs, atom/bond features, 12 SOC toxicity outputs

Pre-trial structure prediction

1b. P2: Allometric scaling + random forest, multi-species integration, 150-compound training

Preclinical-to-clinical translation

2a. P3: XGBoost (186 baseline features), 30-day rolling prediction, per-visit update

Individual patient risk scoring

2b. P4: Clinical-BERT fine-tuned on 86K narratives, symptom/severity/concern extraction 42-day median lead time

3a. P5: Gradient-boosted meta-learner, conditional modality weighting, all P1-P4 as inputs

Multi-modal fusion

3b. All: 48 trials, 37K patients, 142K AEs, 4 therapeutic areas, leave-one-area-out CV

Cross-therapeutic evaluation

Part II -- AEPS Scoring Rubric

4a. Prediction Accuracy: AUROC for 30-day SAE occurrence prediction Higher = better

4b. Lead Time: Median days between first positive prediction and AE onset / 60 day target Earlier = higher score

5a. Specificity: PPV at sensitivity 0.80 operating point Higher PPV = fewer false positives

5b. Clinical Actionability: Fraction of correct predictions enabling preventive intervention Time-sufficient for action

6a. Generalisability: AUROC retention on held-out therapeutic area Cross-area transfer

6b. $AEPS = 0.25 * Accuracy + 0.20 * LeadTime + 0.20 * Specificity + 0.15 * Actionability + 0.20 * Generalisability$